UNL DECONVERTER FOR TAMIL

T.Dhanabalan, T.V.Geetha

Resource Centre for Indian Language Technology Solutions (RCILTS) – Tamil, Department of Computer Science and Engineering, Anna University, Chennai, India – 600025 E Mail : { dhanabalan@cs.annauniv.edu, tvgeedir@cs.annauniv.edu }

Abstract *— This paper discusses the interlingua approach to machine translation. Here Universal Networking Language (UNL) has been used as the intermediate representation. In this paper the DeConverter from UNL to Tamil language has been described. The information needed to generate the Tamil sentence is available at different linguistic levels. Tamil being a morphologically rich language allows a large amount of information including syntactic categorization, and thematic case relation to be generated from the morphological level itself. Information about relating concepts like verbs to thematic cases, adjectival components to nouns and adverbial components to verbs is available through syntactic functional grouping that has been done by the specially designed syntactical generator taking into consideration the requirements of the Tamil language.*

1. INTRODUCTION

There are many possible approaches to machine translation. One approach is the interlingual machine translation system where in the Source Language (SL) text is analyzed using source language dictionary and grammatical information and converted into an interlingual representation. This interlingual representation along with SL to Target Language dictionary and grammar is synthesized to generate Target Language text [7]. In this paper the interlingua structure used to translate from source language to target language is the semantically biased language independent UNL representation. Any translation system using UNL as intermediate representation needs to have an EnConverter from the source language to UNL and a DeConverter from UNL to target language.

In this paper the DeConverter from UNL to Tamil language has been described. Tamil is a free word order language basically because of its rich morphological characteristics. The attachment of constituents to other constituents of the sentence is achieved through morphology and grouping across words through syntactic functional attachments and not through fixed word ordering of words but modified to map with UNL structures. Moreover, in addition to individual word formation using morphological features, there is need to syntactically arrange morphologically formed words using the binary relation and other information available from the UNL structures.

English is predominantly a fixed word order language and follow the subject, verb, object (SVO) pattern while Hindi is a partial free word language with subject, object, verb (SOV) structure [8]. In Hindi the free word order is between adjacent words of a syntactic constituent and hence local word grouping plays an important part. Tamil is predominantly a free word order language which again normally but not necessary follows subject, object, verb (SOV) pattern. This free word order nature of the language is possible due to the thematic cases of noun and person, number and gender markers in the case of verbs being conveyed by morphological suffixes. In this work large part of the information available in the UNL structure is tackled by the morphological generator module of the DeConverter.

2. UNIVERSAL NETWORKING LANGUAGE (UNL)

2.1 Introduction

The Universal Networking Language (UNL) is an electronic language in the form of a semantic network that act as an intermediate representation to express and exchange every kind of information [2].

The UNL represents information, i.e. meaning, sentence-by-sentence. Sentence information is represented as a hyper-graph having Universal Words (UWs) as nodes and relations as arcs. This hyper-graph is also represented by a set of directed binary relations between two of the UWs present in the sentence. Nodes, or Universal Words (UWs) are words based on English and disambiguated by their positioning in a knowledge base (KB) of conceptual hierarchies [1]. Function words, such as determiners and auxiliaries are represented in the form of attributes to UWs, provided that these function words contribute to the meaning.

Binary relations are the building blocks of UNL sentences. They are made up of a relation and two UWs. Relations that link UWs are labeled with semantic roles of the type such as *agent, object, experiencer, time, place, cause*, which characterize the relationships between the concepts participating in the events or states a natural language sentence. UNL has specified forty such relations and claim that these relations are sufficient to represent the interconnection

expressed by natural language sentences. The more details of UNL has been described in a number of papers [2], [3], [5], [6], [9] and [10].

2.2 UNL Features

The text – once converted into UNL – can be converted to many different languages [3]. For example, once a home page is expressed in UNL, it can be read in a variety of natural languages. The meaning representation is directly available for retrieval and indexing mechanisms and tools for automatic summarization and knowledge extraction and it will be converted to a natural language only when communicating with a human user.

UNL greatly reduces the cost of developing knowledge or contents necessary for knowledge processing, by sharing knowledge and contents. Furthermore, if the type of knowledge required for doing some task is described in a language, such as UNL, the software only needs to interpret unambiguous intermediate instructions written in the language to be able to perform its functions.

3. UNL TO TAMIL DECONVERTER

DeConverter is a language independent generator that provides synchronously a framework for word selection, morphological and syntactic generation and natural collocation necessary to form a sentence. DeConverter can convert UNL expressions into a variety of native languages, using a language specific set of Word Dictionary, Grammatical Rules and Co-occurrence Dictionary.

Given a set of UNL structures the primary task is to retrieve the relevant dictionary entries from the Tamil language word dictionary corresponding to the words in the word part of the UNL structures.

The word entries of each language are stored in the Word Dictionary. Each entry of the Word Dictionary is composed of three kinds of elements: Headword, Universal Word (UW) and Grammatical Attributes. The headword is a notation/surface of word of a natural language. UW expresses the meaning of the word, which is to be used as a trigger or link for obtaining equivalent words or expressions. Grammatical Attributes are the information about the behavior of the word in a sentence, which is to be used in deconversion rules.

In certain cases, it is not possible to unambiguously decide on the correct root word for specific UNL word. In such cases there is a need to use a co occurrence dictionary, which specifies more semantic information to solve the ambiguity in the choices of the natural language word. These words are then associated with the relation part of the UNL structures using the UNL identifier tags.

The next step in the DeConversion process is use of language specific, linguistic based deconvertion rules to convert the UNL structure into natural language sentences. These sentences have to obey the morphological and syntactic rules of the language. This is ensured by appropriately building the deconvertion rules which specify the morphological and syntactic structure of the language under consideration.

3.1 Architectural Design of Tamil DeConverter

3.1.1 Semantic Module Architecture

The structure identification module deals with the UNL expression, and it performs the higher-level checks to classify and identify the kind of sentence. After this check, it is forced to assign and compare the verbal root with comparing the aspect, intention, mode and tense. We select and assign the subject of the sentence and again compare with the verbal root, in this case about the number and person. We finally obtain a list of networks, where the roots of these networks (or grammatical trees) are the nuclei of the groups (nominal, adjectival, adverbial,...). This list will be input to the Syntactic Module.

3.1.2 Syntactic Module Architecture

The input for this module is a list of networks (trees) generated at semantic level. The nucleus of each network can be a noun, an adjective, an adverb, a noun affected by a preposition or a relative particle introducing a new clause. The rules of the module have to deal with the position of the adjectives respect to noun, use of pronouns (personal, possessive,...), selection of the adverbs, and so forth.

3.1.3 Morphologic Module Architecture

This module completes the generation of the sentences by adding the endings of the words, specially the verbs (according to the person, number, tense and mode). This module also tackles the insertion of auxiliary verbs for aspect and intention of the sentence. The generation of postpositions, relative pronouns and articles is embedded with the Syntactic module due to the DeConverter constraints. The structure of Tamil DeConverter is shown in FIGURE. 1.

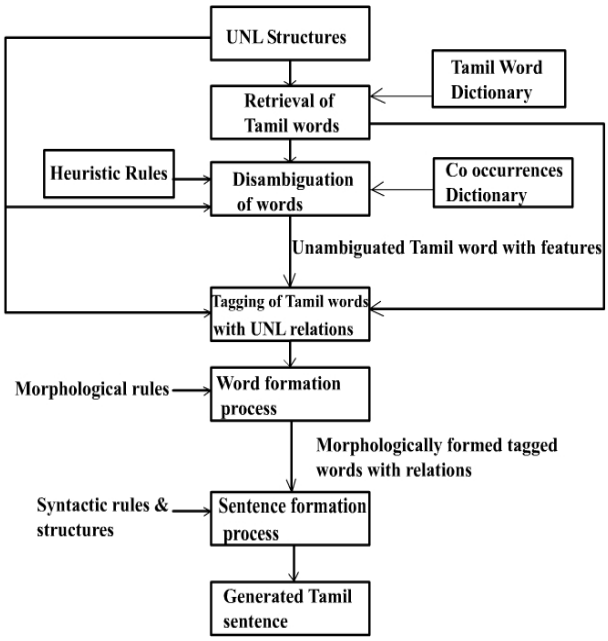

FIGURE. 1
Structure of Tamil DeConverter

The UNL words are retrieved from the UNL structures with the help of UNL – Tamil word dictionary. There may be ambiguity in natural language words. Set of ambiguous with features, UNL relation from the UNL structures and heuristic rules are given to the disambiguation process. This disambiguation process is done with the help of co occurrences dictionary. UNL relations from the UNL structure disambiguated Tamil language words with features and the natural language words with the features are tagged. These tagged UNL words with features and morphological rules are given to the word formation process. Morphologically formed words with the relation and syntactic rules are used for the sentence formation process. This sentence formation process generates the Tamil language sentence.

4. TAMIL LANGUAGE SPECIFIC AND LANGUAGE INDEPENDENT INFORMATION

Basically the DeConverter needs certain information for generating Tamil sentence from UNL input. The information are based on the various linguistic levels like the morphological, syntactic or semantic levels. The amount and type of information needed at each level is largely dependent on the characteristics of the language. Hence the design of the DeConverter is decided by information needed from UNL and the nature of the language, which decides on the type of information that can be generated from the various linguistic levels. UNL has separate concepts for nouns, verbs, adjectives and adverbs in other words there is a need to syntactically categorize the words of the sentence. The attributes to be generated from the concept definition in UNL is number for nouns and tense marker for verbs. The next important part of UNL is the definition of relations. These include case relations of noun concepts with the corresponding verbs, association of adverbial components with verb definitions and the association of adjectival components with the noun definitions.

In this paper the generation of the information for various linguistic levels of Tamil language from the UNL structure has been discussed.

5. MORPHOLOGICALLY GENERATED INFORMATION FOR WORD LEVEL ANALYSIS

Tamil is a morphologically rich language and hence a large amount of information can be generated in the morphological generation phase with the help of analyzing the UNL words and binary relations. The word level Morphological generator for Tamil generates the derivative word for the root word and the various features conveyed by the suffixes. The general noun will have the following syntax.

Noun + (plural marker) + (case marker) + (postposition) + (clitics)

- Example showing number attachment to nouns
  'malar ' (root - flower) + 'kaL' (number - plural) → 'malarkaL' meaning flowers.

- Example showing case attachment to nouns
  'avan' (root - he) + 'ukkaaka' (case - for) → 'avanukkaaka' meaning for him

  'kaTai' (root - store) + 'kku' (case - to) → 'kaTaikku' meaning to store

The morphological generation of verb is much more complex in Tamil. The following is the syntax for verbs.

Verb + (auxiliary verb) + (tense marker) + (PNG marker) + (clitics)

- Example showing verb with simple tense and PNG information
  'cel' (root - go) + 'n' (past tense) + 'aan' (third – person, singular - number, masculine - gender) → 'cenRaan' meaning He went

In most of the languages, prepositions are individual word in a sentence. But in Tamil, there is no preposition and it comes as case marker attached to noun and person, number, gender marker attached to verb. This case ending marker and PNG marker information are used for syntactic functional grouping of words and finding out the correct word for the binary relation. There are several case markers in Tamil like nominative case('o'), accusative case ('ai'),

instrumental case (‘aal’), dative case (‘kku’), locative case (‘il’, ‘iTam’), genitive case (‘in’, ‘uTaiya’, ‘atu’), and ablative case (‘iliruntu’, ‘iTamiruntu’),…
All the PNG marker endings for Tamil are given in the following table.

TABLE 1
PNG marker endings for Tamil

Number	Gender	Person		
		1	2	3
Singular	Masculine			aan
	Femine			aaL
	Mas/Fem	yean	iir,aay	aar
	Neuter	athu, um, a		
Plural	Masculine			
	Famine			
	Mas/Fem	oom	iirkaL	aarkaL, anar, ar, oor
	Neuter	ana		

6. SYNTACTIC GENERATION BASED ON BINARY RELATION AND MORPHOLOGY.

Tamil as already discussed is a relatively free word order language. The free orderness of the language is achieved by two mechanisms namely richness in morphology and cues available that enable syntactic grouping of words. The job of the syntactic generation in the case of Tamil DeConverter is generating Tamil sentence with the help of binary relation and morphological rules. Though the morphological level does provide information about the possible case relations between the main verb and noun components there is ambiguity in some cases. This ambiguity is resolved using a combination of information obtained at the syntactic and semantic levels. In this paper the syntactic generator has been designed to extract the required syntactic and semantic information in order to build the complete sentence.
Let us consider the following UNL expression.

```
[S]
{unl}
[W]
he(icl>person).@generic : 0
very(icl>concept).@generic : 1
run(icl>do).@past.@entry : 2
fast(aoj>thing) : 3
 [/W]
[R]
2 agt 0
3 man 2
1 mod 3
[/R]
{/unl}
[/S]
```

Tamil words equivalent for the universal words are collected from the word dictionary. ‘avan’, ‘mika’, ‘ooTu’, ‘vekam’ are the Tamil word equivalents for the universal words ‘he’, ‘very’, ‘run’, and ‘fast’ respectively. From the binary relation in UNL, ‘avan’ is agent of the action ‘ooTu” and the verb ‘ooTu’ is mentioned as past tense in UNL format. If the agent is third person singular means, the gender marker ‘aan’ is added to the verb ‘ooTu’ and the past tense marker ‘n’ is also added. ‘vekam’ defines the way to carry out an event ‘ooTu’ and it is adverb. So ‘aaka’ is added to the adverb ‘vekam’. The word ‘mika’ is intensifier and it modifies the adverb ‘vekam’ and it is simply added before adverb ‘vekam’. So the generated sentence from the above information is

‘ avan mika veekamaaka ooTinaan’.

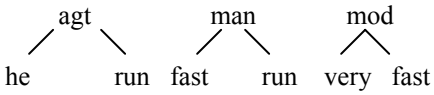

FIGURE 2. Binary relation in UNL format

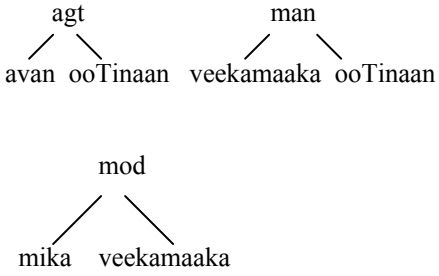

FIGURE 3. Binary relation for Tamil words

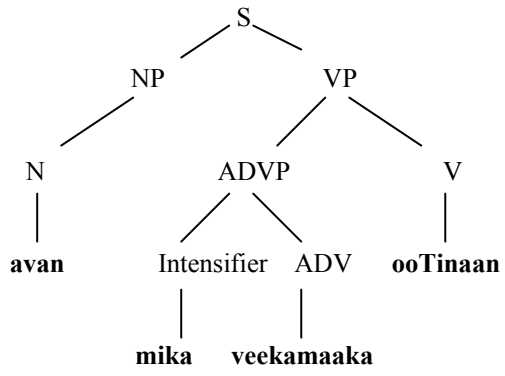

FIGURE 4. Syntactical structure of the sentence

Endings in Tamil words and binary relations are used to generate a sentence. Endings with syntactic functional grouping and semantic information are used for generating correct sentences. The following table

shows some binary relation and its definition and associated endings in Tamil words.

TABLE 2
Binary relation, Definition and its Tamil endings

Binary Relation	Definition	Endings or words in Tamil
and	conjunctive relation between concepts	“ maRRum “ “meelum”
aoj	thing which is in a state or has an attribute	" aaka "
bas	thing used as the basis for expressing degree	“ poola” “kaaTTilum” “vita” “ “mika”
ben	not directly related beneficiary	" ukkaaka " " ukku "
cag	thing not in focus which initiates an implicit event which is done in parallel	" uTan" and " ooTu "
cnt	equivalent concept	“ enpathu “, “ enappaTuvathu”
cob	thing that is directly affected by an implicit event done in parallel	“ uTan “ & may ends with “um”
coo	co occurred event or state for a focused event or state	" konTu "
dur	a period of time during an event occur	" pootu "
fmt	range between two things	" iliruntu " " mutal " " varai "
frm	origin of a thing	" iliruntu "
gol	the final state of object .	" ukku "
ins	instrument to carry out an event	" aal " " uTan "
man	way to carry out an event	Adjective/adverb like “ mika “ “ mikavum”
mod	thing which restrict a focused thing	" aaka "
obj	thing in focus which is directly affected by an event	adverb ends with “a”
opl	place in focus where an event affects	" il " " miitu "
or	disjunctive relation between two concepts	“ allathu “
plc	place of an event occurs	" il "
plf	place of an event begins	" iliruntu " " mutal "
pof	concept of which a focused thing is a part	" in " " uTaiya "
pos	possessor of a thing	" uTaiya "
ptn	non focused initiator of an action	" uTan "
pur	purpose of an event	" ukkaaka " " kku "
qua	quantity of a thing or unit	" niRaiya", atikamaana , “kuRaivaana "
rsn	reason that an event happens	“ Enenil “
seq	prior event or state of a focused event	“ munnal” “ pinnaal” & ends with “kku”
src	initial state of object or an event	“ilirunthu “
tmf	time of an event that starts	" iliruntu " " mutal "
tmt	time of an event that ends	" varai "
to	destination of a thing	" kku"
via	intermediate place or state of an event	" vaziyaaka" , "vaziyee"

7. CONCLUSION

In this paper the development of Tamil DeConverter is described. In Tamil most information for generating sentence from UNL structure is tackled in morphological and syntactical level. In case of ambiguity, there is need to go for complete semantic processing. Relation table is used to find out the words or endings for the specified binary relation. This table information plays important role in both Tamil EnConverter and Tamil DeConverter. The use of UNL

as intermediate representation makes translation of Tamil language available worldwide using a standardized format.

REFERENCES

[1] Munpyo Hong & Oliver Streiter (1999): *Overcoming the Language Barriers in the Web: The UNL-Approach*, 11. Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung (GLDV'99), Frankfurt am Main, Deutschland.

[2] Hiroshi Uchida, Meiying Zhu, Tarcisio Della Senta (1999) : *A gift for a Millennium.* The United Nations University.

[3] UNL Center (2000) : *Enconverter Specification*, UNDL Foundation.

[4] UNL Center (2000) : *DeConverter Specification*, UNDL Foundation.

[5] Serrasset, G. and Boitet, C. (2000). *On UNL as the Future "html of the linguistic content" & the Reuse of Existing NLP Components in UNL-related Applications with the Example of a UNL-French Deconverter.* Proceedings of the 18 th International Conference on Computational Linguistics, pp. 76□771.

[6] Bouguslavsky, I., Frid, N. and Iomdin, L. (2000). *Creating a Universal Networking Module within an Advanced NLP System*. Proceedings of the 18 International Conference on Computational Linguistics, pp. 83-89.

[7] Arnold, D. et al. (1994) *Machine translation: an introductory guide.* Manchester/Oxford: NCC/Blackwell.

[8] Shachi Dave, Jignashu Parikh and Pushpak Bhattacharyya, *Interlingua Based English Hindi Machine Translation and Language Divergence* , to appear in Journal of Machine Translation, vol 17, 2002.

[9] Pushpak Bhattacharyya, *Many Languages on the Net*, PCQuest Magazine, September 2002.

[10] P. Bhattacharyya, *Knowledge Extraction into Universal Networking Language Expressions*, in Universal Networking Language Workshop, Geneva, Switzerland, January, 2001.